图书情报工作
第63卷 第5期 2019年3月

知识直播平台付费用户群体画像研究

■ 张莉曼 张向先 卢恒 张玲
吉林大学管理学院 长春 130022

摘要：［目的/意义］为知识直播平台精细化地定位人群特征、进行针对性运营并挖掘潜在价值用户提供有效的思路和方法。［方法/过程］以知识直播平台为研究对象，首先设计群体画像的概念模型，然后提出基于密度峰值聚类的知识直播平台付费用户群体画像的方法，最后以知乎 live 平台的付费用户数据为样本，使用 Matlab 中的密度峰值聚类算法对样本数据进行群体划分。［结果/结论］实验结果表明该平台存在5类典型的用户群体，通过对聚类中心点的对比分析，识别出各群体典型特征，并提出精准运营策略。

关键词：知识直播平台 密度峰值聚类 群体画像 付费用户
分类号：G312
DOI：10.13266/j.issn.0252-3116.2019.05.010

互联网络信息中心（CNNIC）发布的第42次《中国互联网络发展状况统计报告》数据显示，截至2018年6月底，网络直播用户数量达到4.25亿，占网民总数的53%[1]。在这种大背景下，以知识即时传播为目的的知识直播平台异军突起，现已成为用户进行知识交互的新型社交平台。与此同时，共享经济催生了知识直播平台中付费产品的上线，用户通过支付一定费用，按自身需求获取针对性的优质内容。用户持续付费是知识直播平台实现商业价值的关键，因此，对付费用户进行群体画像，识别其典型特征及潜在需求，对于平台提供精准服务、维系核心用户与制定发展战略至关重要。

用户画像作为勾画目标用户、提高决策效率的有效工具，现已在各领域得到广泛应用。在电子商务领域，K. Sugiyam 等抓取了用户 览行为、结果评价等信息，构建用户的偏好模型并进行针对性推送，实现用户与平台的协同交互[2]。在社交网络领域，王凌霄等从用户的资历、参与度、回答质量、发展趋势4个方面构建了社会化问答社区的用户画像，并采集知乎数据验证结果，为进行社会化问答社区的用户分层、行为监控提供参考[3]；林燕霞基于社会认同理论按照主题偏好对微博用户群体分类，并利用多文本挖掘其主题偏好，通过用户特征属性的提取，将微博用户归纳为5类群体[4]。在移动通信领域，丁伟等在保证用户隐私安全的前提下，通过分析用户的通信记录研究用户画像在个人征信中的应用[5]；在图书馆领域，王庆等研究了图书馆用户画像模型，并从单用户与多用户两个层面设计了图书馆馆藏资源的精准推荐模式[6]。由此可见，用户画像已成为学术界关注的热点，但由于知识直播平台尚属新兴事物，目前将用户画像应用于知识直播平台的研究相对较少，对知识付费背景下付费用户的特征及群体画像的研究更为匮乏。

因此，本文立足于国内外研究现状，将用户画像思想应用于知识直播平台的付费用户群体。通过构建付费用户群体画像概念模型，采集平台上的客观数据并进行聚类分析，把平台上的付费用户描绘成几类典型用户。将传统的单个用户追踪转变为对群体的抽象概括，从而为平台运营者实现精准化运营、助力产品快速变现提供借鉴参考。

1 相关概念与文献

1.1 知识直播平台

作为知识的生产、传播、消费一体化平台，知识直播平台是指掌握某领域专业知识或技能特长的知识供给方，利用互联网及多媒体技术，将自身的知识技能实时传递给知识需求方，使知识得以有效传播并可进行

作者简介：张莉曼（ORCID：0000-0002-0770-3708），博士研究生，E-mail：326671265@qq.com；张向先（ORCID：0000-0003-3186-2677），教授，博士，博士生导师；卢恒（ORCID：0000-0002-6680-5915），硕士研究生；张玲（ORCID：0000-0002-6654-222x），博士，硕士生导师。

收稿日期：2018-07-17 修回日期：2018-10-14 本文起止页码：84-91 本文责任编辑：王传清

变现的互动性社交平台[7]。目前国内影响力较大的知识直播平台包括知乎 live、美时美刻、知深、荔枝微课等;国外有 Open English、Platzi、Livecoding 等。

当前学者们对知识直播平台的研究多在运营策略、传播模式、用户行为等方面,例如杨静运用长尾理论分析了知乎 live 平台知识生产、传播、消费三阶段的运营模式[8];王亮研究了实时语音直播社区的传播模式与知识变现路径[9];用户行为层面多关注用户参与动机[10]或用户付费影响因素[11]。对知识直播平台用户画像的应用仅为用户分层的理论性分析[7],缺乏以大量数据为支撑的实践性探索。

1.2 付费用户群体画像

用户画像也称用户标签化,即根据用户的社会属性、生活习惯、消费行为等数据标签,抽象出一个能够概括用户全貌的模型[12]。用户画像包括单个用户画像与用户群体画像两个方面,单个用户画像要求用尽量多的标签描述一个用户的特征,目的是展示不同用户个体的独立性与差异性,但不利于分析大规模用户数据或制定群组推荐策略[13]。群体画像是指运用分类、聚类等方法根据标签数据计算用户间的相似度并把具有相似特点的用户划分到同一类簇后再进行特征描述[6]。引入群体画像思想将平台上的付费用户归纳为几个典型的虚拟人物,抽象出共性特征并描绘出具有差异化的用户模型,目的是为平台运营者在大数据环境下快速把握付费用户特征、制定精准营销战略提供参考。

当前群体画像已在网站、图书馆、学科服务等领域得到有益尝试。J. A. Iglesias 等挖掘了网页上的日志数据并运用聚类方法刻画出不同群体的用户画像[14];王顺箐基于社会网络的聚合策略进行了图书馆群体推荐系统的可行性探究[15];薛欢雪从 4 个维度构建了学科服务用户画像模型[16]。上述研究为付费用户群体画像设计提供了一定的思路,但这些研究多为理论性的宏观画像,不利于处理大规模用户数据,且画像粒度较为粗糙。由于知识直播平台在提供付费知识的同时,还兼具社交功能,付费用户群体的标签构成更为复杂,数据量也更大,因此在构建付费用户群体画像时应综合考虑付费用户的多方面属性及特殊性,采集大量客观数据,构建立体、细致、针对性强的群体画像。

1.3 知识直播平台付费用户群体画像方法的提出

构建用户群体画像是聚类算法的典型应用场景之一[17],通过聚类得到差异化群体模型,可在大数据环境下高效分析大规模用户数据,把握用户特征及需求[18]。这一思想已得到广泛应用:吴江等使用 k-means 算法对用户聚类来识别用户角色[19];杨卫红等使用 k-means 算法对用户的用电行为进行聚类[20]。但是 K-means 聚类需人为事先设定类簇个数,而类簇的选定往往难以估计。陈娟等[21]使用层次聚类来产生初始聚类中心,E. A. Mohammed 等提出 EIAgha 初始化算法,认为可以根据数据的排列形状确定聚类中心点[22],但上述方法计算复杂、只能得到局部最优结果,不利于处理大规模复杂数据。

密度峰值聚类算法(Density Peaks Clustering Algorithm,DPCA)是由 A. Rodriguez 和 A. Laio 提出的一种基于密度和距离的聚类算法[23]。它能够快速发现任意形状、规模的类簇中心,不需人为确定类簇个数,样本点归类无需迭代求解,适用于大规模的数据处理,已经应用于遥感图像分析、社交网络、文本发现、文本摘要、图像分类等多个领域[24]。知识直播平台的付费用户数据量大、数据分布形态未知,无法事先确定合理的类簇个数。因此本文选用密度峰值算法对样本进行聚类,充分发挥密度峰值聚类在处理高维数据时的优越性,完全依靠无监督的聚类结果,保证群体画像的客观性。

2 知识直播平台付费用户群体画像设计

2.1 知识直播平台付费用户群体画像构建流程

从用户群体画像的概念[6]可以看出,构建群体画像包含用户标签确定、数据采集处理、方法选择与实验、画像呈现等环节。典型的用户画像构建方法有 A. Cooper 的"七步人物角色法"和 L. Nielsen 的"十步人物角色法"[25],这两个方法在流程上可概括为研究并获取用户信息、细分用户群、建立并丰富用户画像三个阶段。因此,本文结合用户群体画像概念及"七步""十步"人物角色法,确定知识直播平台付费用户群体画像的构建流程:①设计概念模型。结合用户的静态数据与动态数据,从多个维度设计付费用户画像标签。②运用密度峰值聚类算法实现群体划分。过程包括数据采集、预处理、变量确定与仿真实验。③根据聚类结果提取类别特征,呈现群体画像并提出精准运营策略。见图 1。

2.2 知识直播平台付费用户群体画像的概念模型

用户群体画像具有较强的领域性,构建时应充分考虑实际应用场景,反映出情景化用户特征。知识直播平台的付费用户群体具有一定的特殊性。首先,当

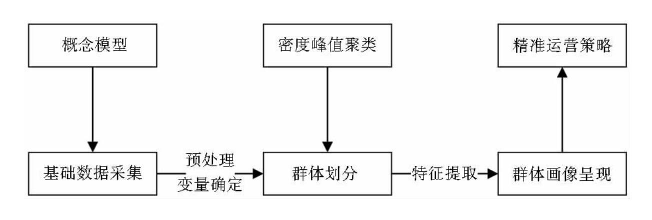


图 1 知识直播平台付费用户群体画像构建流程

前的知识直播平台多为综合型,涵盖主题丰富,因此其付费用户来源广泛、专业需求突出。其次,相较于传统的教育类培训平台,知识直播平台不仅是传递知识的载体,也提供关注、提问的交流功能,因而其付费用户互动性强、行为偏好标签更为丰富。最后,付费用户群体通过支付费用获取知识,具有一定的经济基础,消费特征明显。上述特点使得付费用户群体的标签构成更为复杂,因此本文在构建付费用户群体画像时综合考虑付费用户的多维特征,充分挖掘用户静态的基本属性及动态的行为属性,从付费用户基本属性、付费用户在平台关注或提问而形成的偏好属性以及付费用户在购买决策过程中的价值属性 3 个维度构建知识直播平台付费用户群体画像概念模型。

2.2.1 付费用户基本属性 付费用户基本属性由结构化的静态数据组成,反映了付费用户的基本特征,是构建用户画像的基础。其中,性别特征是群体行为、偏好以及需求趋向的影响因素之一,因此构建付费用户群体画像需识别性别特征。由于知识直播平台以知识的定向传播为目的,用户根据自身兴趣及需求参与相应的直播课程,因此在构建画像时应考虑用户的学历、专业、所在行业、企业及职位。此外,居住地体现了付费用户的地域特征。因此,付费用户基本属性由性别、学历、专业、行业、企业、职位、居住地 7 个标签组成。

2.2.2 付费用户偏好属性 付费用户偏好是指付费用户对某一事物的倾向及关注程度,用户往往通过日常的行为偏好来辅助购买决策。知识直播平台是以内容为主导的价值型平台,有利于用户高效筛选信息,满足对特定知识的需求。因此,在确定用户的偏好属性时,应考虑用户的直观需求及关注倾向。以即时问答类知识直播平台——知乎 live 为例,用户可以通过"提问"功能直接表达自己的知识诉求,以"收藏"的方式进行知识沉淀,满足自身的个性化需求。同时,关注感兴趣的讲师及其他用户、关注相关话题内容等体现出用户在平台上的活跃程度与意愿倾向。从这一维度考虑,付费用户偏好属性应包括提问数、收藏数、关注人数、关注内容数 4 个标签,其中提问数属于显性需求标签,其余 3 个属于隐性需求标签。

2.2.3 付费用户价值属性 付费用户价值属性旨在判断用户对于平台的价值,即付费用户对平台的利润贡献情况。引导普通用户完成付费并持续付费,提高用户忠诚度及消费黏度,不断为平台输出价值,是知识直播平台实现知识变现的主要途径。因此,了解用户的价值体现,挖掘高价值群体并针对性地刺激消费是平台运营发展的关键。用于用户价值分析的 RFM 模型从最近消费时间(recency)、消费频率(frequency)、消费金额(monetary)3 个方面细分了用户群体与用户价值[26],除此之外,分析用户首次消费时间及参与次数对于及时挖掘新付费用户、维系忠诚用户群体、促进用户持续付费具有重要意义,因此这一属性维度选取了首次消费时间、末次消费时间、参与次数、消费频率、消费金额 5 个标签。

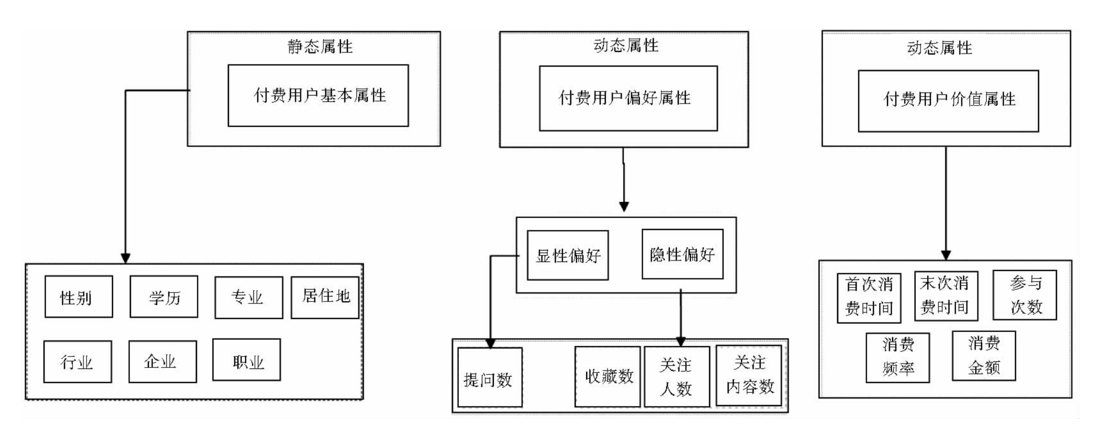


图 2 付费用户群体画像概念模型

3 基于密度峰值聚类的知识直播平台付费用户群体画像的实现

3.1 密度峰值聚类原理及步骤

3.1.1 密度峰值聚类原理 密度峰值聚类算法有两个基础假设[27]：①聚类中心始终被低密度点包围；②聚类中心点之间相距较远。对于每个数据点 x_i，都存在一个局部密度值 ρ_i，当数据点为离散时，利用公式(1)计算局部密度值 ρ_i；当数据点为连续时，采用公式(2)来计算局部密度值 ρ_i。

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \qquad 公式(1)$$

$$\rho_i = \sum_j \exp\left(-\frac{s_{ij}^2}{d_c^2}\right) \qquad 公式(2)$$

其中函数

$$\chi(x) = \begin{cases} 1,\ x < 0 \\ 0,\ x \geqslant 0 \end{cases} \qquad 公式(3)$$

DPCA 算法采用欧式距离计算数据点两两之间的距离，假设存在数据点 x_i 和 x_j，则它们之间的距离公式为：

$$dist(x_i, x_j) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad 公式(4)$$

d_c 表示截断距离（cutoff distance），其取值影响 DPCA 聚类的结果，该参数一般由升序排序的数据点距离中的前 1%－2% 的距离所决定。

对于每个数据点 x_i，还存在一个距离 δ_i，其计算公式如下：

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} d_{ij},\ i \geqslant 1 \\ \max(d_{ij}),\ \rho_{ij}为全局最高 \end{cases} \qquad 公式(5)$$

当数据点 x_i 具有最大局部密度时，δ_i 表示数据集 S 中与 x_i 距离最大的数据点到 x_i 的距离；否则，δ_i 表示在所有的局部密度大于 x_i 的数据点中，与 x_i 距离最小的那个数据点到 x_i 的距离。

3.1.2 密度峰值聚类步骤 DPCA 的核心思想是同时考虑局部密度值 ρ_i 和距离 δ_i，采用二维平面提取聚类中心点，即把局部密度值 ρ_i 看做横轴，把距离 δ_i 看做纵轴，提取密度较大且与其他参考点之间距离较大的点作为聚类中心点[23]。DPCA 算法具体步骤如下：

步骤 1：计算截断距离 d_c。首先根据公式 4 计算数据点间距离 d_{ij}；然后将数据点间距离 d_{ij} 按升序排序；最后根据排序选择合适的数据点来决定 d_c 的值。

步骤 2：选取类簇中心。首先，根据公式(1)、公式(2)和公式(5)分别计算 ρ_i 和 δ_i；然后将所有备选点的密度按降序排序；最后把具有高 ρ_i 值和相对较高 δ_i 值的备选点标记为类簇中心点。

步骤 3：分配非类簇中心点到相应的类簇中心。即根据 δ_i 值的从属关系将非类簇中心点依次依附于其更高密度中心点上。

步骤 4：聚类结果展示。

3.2 数据采集与预处理

3.2.1 数据采集 本文以即时问答类知识直播平台——知乎 live 为研究样本。在知乎 live 平台中，主讲人首先创建一个 live 直播群，然后系统自动将 live 推送给关注主讲人的用户，用户点击并支付相应金额（由主讲人设定）便可进入到直播群中。直播群以语音形式分享专业有趣的知识，并通过即时互动提高信息交流效率。为了保证数据的客观性和代表性，本文选取 2016 年 6 月－2018 年 1 月间的 8 场不同类别的 live 为数据样本，涉及职业、教育、艺术、互联网、心理学、音乐影视、医学健康与生活方式 8 个领域。按照付费用户群体画像概念模型提出的 16 个标签，利用八爪鱼采集器批量采集了 18 520 个付费用户的相关信息，共 287 523 个数据。将获取的付费用户数据存储于 MySQL 数据库中，作为知识直播平台付费用户群体画像的样本数据。

3.2.2 数据预处理 用户注册信息不完整、不规范以及数据形式多样会影响聚类效果。在运用聚类算法进行用户群体划分之前，需要对采集的信息进行预处理，主要包括以下几个步骤：

（1）样本数据清洗。首先，通过初步筛选，删除包含重复值与异常值的数据。由于用户填写注册信息时往往具有随意性和主观性，需要人工对用户基本信息进行审查，将基本信息数据残缺与不规范的样本予以剔除。这一过程共剔除 3 320 个样本。

（2）剔除部分变量。空值过多会导致数据稀疏，影响聚类效果。因此需剔除空值率超过标准值的变量。定义数据为空白或不能识别的值（null）为空值，映射在某变量的数据空值率＝该变量为空值的样本数/有效样本总数，设定空值率为 25% 以下为有效变量。经过计算，用户学历、专业、企业、职位 4 个变量的空值率分别为 35%、42%、47%、43%，故将其剔除。

（3）转化与编码。原始数据包括数值型数据与文本型数据两类，为了满足聚类分析对数据类型的要求，需将文本型数据转化为数值型数据。将值域为〔男，女〕的性别变量转化为〔1，－1〕；居住地变量下识别出 54 个不同的城市，数据较为分散，不便于分析聚类结果。因此，本文按照《中国城市新分级名单》将 54 个城

市划分到相应层级中,用数字 1 - 4 表示,海外居住地统一归为 0,经过处理后,居住地变量的值域为〔0,1,2,3,4〕。此外,原始数据中首、末次消费时间变量下的数据为时间型数据,为了使数据相对集中便于聚类特征的明显化,将首次消费时间距今 6 个月以内的记为 1,6 - 12 个月记为 2,12 个月以上的记为 3,首次消费时间的值域为〔1,2,3〕;将最近消费时间的标准设定为 90 天,90 天内有消费的记为 1,无消费的记为 0,最近消费时间的值域为〔1,0〕。

(4)数据计算。由于知乎 live 平台提供的公开数据中不包括用户消费金额、消费频率两个变量的数据,但可采集到用户参与的 live 总数以及每场 live 单价,所以对于付费用户价值属性中的消费金额、消费频率(每月)两个变量标签可通过人工计算得出。

经过上述处理后,保留 12 个变量,再次剔除在这些变量的映射中不完整的样本,最终获得有效样本 n =12 545。

3.3 基于密度峰值聚类的知识直播平台付费用户群体画像的实现

3.3.1 归一化处理 本文借助于 Matlab 软件实现聚类实验。原始数据的值域相差较大会影响聚类结果的准确性,因此利用 Matlab 工具箱中的函数进行权值和阈值的初始化,并采用 PREMNMX 函数对样本数据进行归一化处理,以提高聚类算法的收敛效率。

3.3.2 d_c 的确定及聚类中心选择 DPCA 算法根据若干数据集的经验值选取截断距离 d_c,其计算公式为:

$$d_c = D_{sort}(N * percent) \qquad \text{公式(6)}$$

其中,D_{sort} 为数据集的数据点间距离升序集合,N 为该集合的总量,*percent* 为平均邻近点数目百分比。DPCA 算法提出时作者建议 *percent* 一般取值为 1% - 2%[23]。*percent* 取值过大会提高类簇边缘点的局部密度,使类簇过少甚至所有数据点都归到一个类之中。取值过小会导致类簇密度峰值点的核心密度值不够凸显,影响类簇中心点的确定[28]。因此,其最终取值应与研究实际相关。本研究分别选取 *percent* 为 1% - 5% 的多个截断距离,使用 Matlab 软件进行对照实验,限于篇幅,只展示部分数据(见表 1)。

表 1 中,Elements(max)表示类簇中最多的数据点个数,Elements(min)表示类簇中最少的数据点个数,数据点过多过少都会影响聚类结果的代表性与说服力。Silhouette 系数(轮廓系数)是一种衡量聚类结果的指标,其取值在[-1,1]之间,该值越接近于 1,说明簇内越紧凑,离其他簇越远[29]。其计算公式为:

表 1 不同百分比下对照实验结果数据(部分)

percent	centers	Elements (max)	Elements (min)	silhouette	d_c
1%	7	3 885	609	0.337 4	0.096 7
1.5%	6	3 837	667	0.41 76	0.118 1
2.5%	6	5 990	637	0.469 6	0.154 6
3.5%	5	5 936	708	0.555 2	0.183 7
4.5%	5	5 903	771	0.564 7	0.209 5
5%	5	5927	751	0.560 4	0.221 3

$$s(i) = \frac{d_{b(i)} - d_{a(i)}}{\max\{d_{b(i)}, d_{a(b)}\}} \qquad \text{公式(7)}$$

其中,$d_{b(i)}$ 表示数据点 i 与其他类最低平均不相似度,$d_{a(i)}$ 表示数据点 i 与它所在类的平均不相似度。经过实验可知,当 *percent* 取值为 4.5% 时,silhouette 系数最大,且此时每一类簇中数据点数量合理,可对样本数据进行科学合理的群体划分。因此,确定 d_c 为 0.209 5,得到密度峰值决策图(见图 3)。其中横轴 ρ 代表不同类型样本点之间的欧氏距离,纵轴 δ 代表同一类型样本点之间的欧氏距离,即同类型样本的紧密程度[30]。图中可看出明显区分的中心点有 5 个,即样本被分为 5 类。

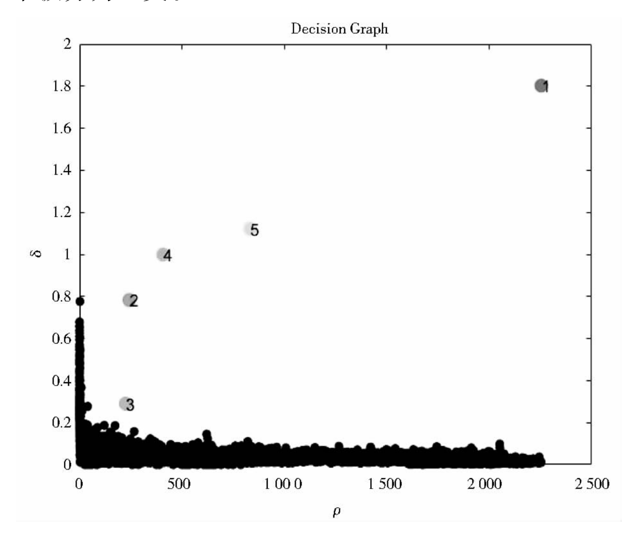


图 3 密度峰值决策图

聚类中心选取效果图(gamma 曲线)如图 4 所示,横轴代表样本编号 n,纵轴代表密度 ρ 与距离 δ 的乘积 γ。gamma 曲线呈下降趋势,并且逐步逼近横轴,说明聚类中心突出,曲线趋于平缓之前的点均可作为聚类中心点。图中曲线平缓之前出现 5 个点,与本文选取的中心点个数一致,证明聚类中心选取得合理。

4 实验结果与分析

4.1 群体划分结果

经过密度峰值算法的聚类,得到 5 个聚类中心点,

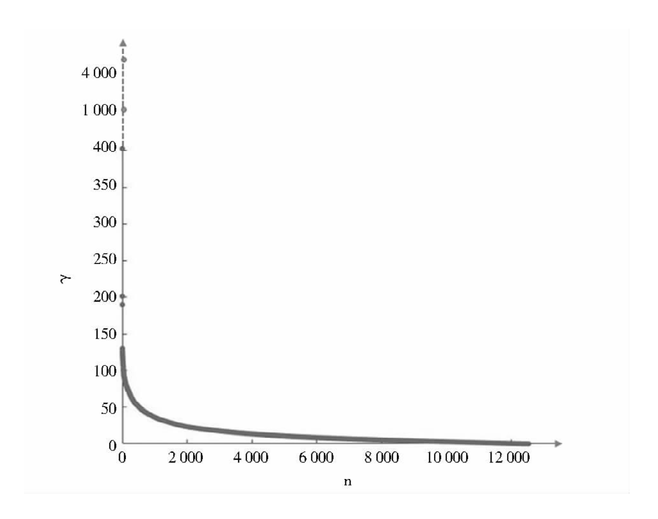


图 4　聚类中心选取效果图

即知乎 live 平台的付费用户样本被划分成 5 类群体，分群结果如表 2 所示。根据表 2 可知，各类别的样本数分别为 5 903、771、804、1 428、3 639，所占比重分别为 47.05%、6.15%、6.41%、11.38%、29.01%。样本数据中的第 706、5 717、8 337、8 637、10 242 组数据为聚类中心点。

表 2　知乎 live 付费用户分群结果

Clusters	Elements	Proportion	Center
1	5 903	47.05%	706
2	771	6.15%	5 717
3	804	6.41%	8 337
4	1 428	11.38%	8 679
5	3 639	29.01%	10 242

4.2　整体画像呈现

为了给用户群体画像分析提供参考标准，从而更好地识别各群体用户的差异化特征，在进行群体画像之前首先对全样本数据进行分析，运用 SPSS20.0 统计软件对样本数据进行描述性统计性分析，刻画知乎 live 平台付费用户整体画像。

整体画像为：男性用户有 8 224 人，占比 65.6%，女性用户有 4 321 人，占比 34.4%，说明知乎 live 平台的付费用户以男性为主。一线城市用户高达 55.1%，二线城市用户占比 24.7%，说明目前知乎 live 平台的主体市场为国内一、二线城市。用户所在行业前 5 位分别是：互联网、金融、教育、计算机、法律，说明付费用户可能主要来源于一、二线城市白领以及在校大学生等对专业技能、自我提升有需求的人群。用户偏好属性维度，显性需求标签提问数小于 10 的占比 77.9%，而隐性需求标签收藏数、关注人数、关注内容数主要集中在 30－50 区间，说明这些用户不善于主动提问而直接暴露需求，因此在运营时应注意把握用户的隐性需求，对于高隐性需求的用户可以通过回访、调研等方式明确用户偏好，通过推送相关课程触达用户，以引起用户的需求共鸣。价值维度方面，消费金额在 1 000 元以上的仅占 11.9%，主要集中在 50 元以下和 100 元－400 元，消费水平以中低等为主。

4.3　群体画像呈现及精准运营策略

聚类中心点即密度较高、相互距离较远的点，各类中其他的点均以此为中心，因此可将密度中心点视为各群体的典型代表。通过类比各类簇中心点（见表 3），结合各个标签变量的取值范围，可以分析得出各群体用户差异化的属性特征，得到知识直播平台付费用户的群体画像。

表 3　各类簇样本中心点

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
性别	1	1	1	1	1
居住地	1	2	2	2	3
行业	1（互联网）	4（法律）	2（教育）	6（金融）	2（教育）
收藏数	70	18	4	79	1
关注人数	180	55	300	182	17
关注内容数	72	5	5	55	5
提问数	37	1	2	38	0
参与次数	79	7	10	4	2
首次消费	3	3	2	2	1
末次消费	1	0	1	0	1
消费频率	5.64	0.70	3	1.33	1
消费金额	1975	122.5	99	79.2	39.8

经过对比分析可得到 5 类用户的群体画像：

4.3.1　忠诚型价值用户　所占比例将近一半，这类群体中男性居多，多数在北、上、广、深等一线城市从事互联网、金融等行业，说明知乎 live 在一线城市中拥有坚实的用户基础。此类用户偏好属性较为明显，关注了较多的知识直播课程主讲人及相关话题，善于通过收藏进行知识沉淀，也会在平台上积极提问。价值属性方面，此类用户购买能力强，消费金额多在 1 900 元左右，平均每月消费 5－6 次，使用平台的时间较长，并且多数用户近期仍有消费行为，少数无消费的用户参与次数及消费频率均较高，证明与平台的粘性较好，具有一定的品牌忠诚度。该群体是知识直播平台得以持续发展的关键，因此应进一步提高这些用户对平台的认可度，不断优化用户体验，可以通过积分、等级、特权等方式提高用户的身份价值，同时激励这些用户积极参与平台内容贡献，实现从被动获取知识到主动输出优质内容的转化，实现平台的良性循环。

4.3.2　沉睡型流失用户　此类用户人数较少，仅占 6.15%。其中男性偏多，多居住于二线城市，法律、医疗等专业性较强的行业较多。该群体的典型特征是首次

消费时间为12个月以前,属于老客户群体,但是最近90天无消费行为,并且往期消费频率低,平均每月消费不到一次,说明这一类别的用户属于不活跃群体并且已对平台失去兴趣。然而也应注意到,该类群体往期消费金额较多,导致流失情况可能因为用户体验不尽人意,碎片化知识无法满足系统学习的需求。也可能与用户所从事的行业有关,用户往往为了解决某一具体问题而付费,持续性消费意愿不强烈。鉴于此,平台运营者可从两方面挽回已流失的用户:一方面,通过回访调研用户流失的原因,充分挖掘其潜在需求,进行平台及内容的更新迭代以优化用户体验;另一方面,通过消息提醒、课程推送等方式与用户建立联系,激活用户再次使用的意愿。

4.3.3 社交型经济用户　占比较少,男性偏多。这类群体大多居住在二线城市,所在行业多为教育或者计算机。该群体关注的人数在5类群体中最多,体现出明显的社交属性,但是收藏、关注话题内容、提问的数量较少,用户需求特征并不明显。价值属性方面,首次消费时间在3－12个月以前,近期仍有消费,说明用户与平台黏度较高。同时消费次数较多、频率也比较高,属于活跃性用户群体,但是所消费金额并不多,说明该类用户在进行知识付费时主要选择单价较低或者有折扣优惠的课程。该群体用户往往没有急切性的需求,付费意愿易受到价格影响,从而在平台上进行稳定、持续性的消费投资。因此运营过程中应善于把握该类用户特点,通过发放优惠券、推送折扣信息等方式刺激其不断消费。

4.3.4 需求型潜力用户　这类群体占比为11.38%,主要为二线城市的男性用户,所在行业多为金融、计算机。该群体在偏好属性上表现为高收藏、高关注、高提问,说明其需求偏好特征明显,属于主动学习型的用户,消费潜力大。但是该群体对于平台的价值并不高,具体体现为消费次数少、消费频率较低,首次消费时间距今较远且近期无消费行为,这就说明该群体用户的消费行为并不活跃,对平台的依赖程度低。原因可能在于用户对以往的付费体验不满意,认为知识直播不能满足其个人需求。针对该人群,可以根据他们的偏好属性挖掘出需求主题,实现针对性、个性化的定制推送,改善用户体验,进而激发其消费潜力。

4.3.5 新兴待激励用户　所占比例较大,仅次于忠诚型价值用户群体。男性较多,大部分居住于三线城市,教育领域的用户较多。这类群体的首、末次消费时间均在3个月内,说明他们使用该知识直播平台的时间不长,现阶段仍属于留存用户。偏好属性方面的社交与需求特征均不明显,消费次数、频率、金额也不高,原因可能在于用户刚刚接触这类平台,对平台的信任感、归属感还有待加强,或者对平台的各项功能还不够了解。因此针对此类用户,平台应尽快挖掘并满足其付费需求,用各种运营手段触达用户,告知活动信息,积极引导新兴用户向活跃用户转变。同时建立用户激励体系,在用户提升等级并获得成就感的同时,也深化用户对平台的了解。此外,该类群体主要居住于三线城市,并且人数较多,说明知识直播平台辐射范围已从一、二线城市向三、四线城市扩散,三、四线城市市场前景广阔。

5 结语

本文以知识直播平台的付费用户为研究对象,首先从付费用户基本属性、偏好属性、价值属性3个维度提取出16个标签。然后利用八爪鱼采集器采集了知乎live平台18 520组付费用户数据,经过预处理之后保留12 545组数据作为样本,再运用密度峰值聚类算法将样本数据划分为5个群体。最后通过类比各群体的聚类中心点并与整体画像对比,呈现出知乎live平台付费用户的5类群体画像:①忠诚型价值用户,消费金额及频率高,对平台价值大;②沉睡型流失用户,近期无消费行为,在平台表现不活跃;③社交型经济用户,社交属性明显,通常只会为单价低的产品付费;④需求型潜力用户,需求偏好特征明显,但是付费并不多,消费潜力大;⑤新兴待激励用户,近期刚刚开始使用平台,需要平台激励从而向活跃用户转化。

本文构建的基于密度峰值聚类的知识直播平台付费用户群体画像应用性较强。首先,采用密度峰值聚类算法进行群体划分,充分发挥了密度峰值聚类在处理多维数据时的优越性,聚类速度大大提高;其次,不需人为事先确定群体个数,聚类结果更为合理。知识直播平台可根据付费群体画像挖掘其特征及需求,从而进行精准运营,针对性地完善平台产品及服务,助力产品快速变现。

由于平台数据的保密性,本文利用爬虫抓取的数据还不够全面,例如缺乏动态性的用户浏览数据、用户搜索数据等,后期研究中将着重考虑如何获取更为全面的标签数据,细化用户画像的颗粒度,刻画更为精准、丰富的用户画像。

参考文献:

[1] 中国互联网信息中心. 第42次中国互联网络发展状况统计报告[EB/OL]. [2018－08－20]. https://tech.sina.com.cn/i/2018-08-20/doc-ihhxaafz2145674.shtml.

[2] SUGIYAMA K, HATANO K, YOSHIKAWA M. Adaptive web search based on user profile constructed without any effort from users[C]//Proceeding of the 13th international conference on World Wide Web. New York: ACM, 2004:675－684.
[3] 王凌霄,沈卓,李艳. 社会化问答社区用户画像构建[J]. 情报理论与实践,2018,41(1):129－134.
[4] 林燕霞,谢湘生. 基于社会认同理论的微博群体用户画像[J]. 情报理论与实践,2018:1－11.
[5] 丁伟,王题,刘新海等. 基于大数据技术的手机用户画像与征信研究[J]. 邮电设计技术,2016(3):64－69.
[6] 王庆,赵发珍. 基于"用户画像"的图书馆资源推荐模式设计与分析[J]. 现代情报,2018(3):105－109,137.
[7] 赵鑫,赵盼超. 国内外知识直播平台内容创业模式比较研究[J]. 中国出版,2017(23):15－20.
[8] 杨静. 长尾理论视阈的知识分享变现条件分析[D]. 南京:南京大学,2017.
[9] 王亮. 基于语音互动的知识付费问答社区的传播模式研究[D]. 上海:上海师范大学,2018.
[10] 张结红. 知识分享类平台用户参与动机对行为的影响研究[D]. 合肥:安徽大学,2018.
[11] 赵杨,袁析妮,李露琪等. 基于社会资本理论的问答平台用户知识付费行为影响因素研究[J]. 图书情报知识,2018(4):15－23.
[12] 杨双亮. 用户画像在内容推送的研究与应用[D]. 北京:北方工业大学,2017.
[13] 何娟. 基于用户个人及群体画像相结合的图书个性化推荐应用研究[EB/OL]. [2018－10－10]. http://www.cnki.com.cn/Article/CJFDTotal-QBLL20180816003.htm.
[14] IGLESIAS J A, ANGELOV P, LEDEZMA A, et al. Creating evolving user behavior profiles automatically [J]. IEEE transactions on knowledge and data egineering, 2012,24(5):854－867
[15] 王顺箐. 以用户画像构建智慧阅读推荐系统[J]. 图书馆学研究,2018(4):92－96.
[16] 薛欢雪. 高校图书馆学科服务用户画像创建过程[J]. 图书馆学研究,2018(13):67－71,82.
[17] 陈蕾夷. 智能化用户分群模型的研究与实现[J]. 电脑知识与技术,2018,14(19):1－3.
[18] 陈添源. 高校移动图书馆用户画像构建实证[J]. 图书情报工作,2018,62(7):38－46.
[19] 吴江,侯绍新,靳萌萌,等. 基于LDA模型特征选择的在线医疗社区文本分类及用户聚类研究[J]. 情报学报,2017,36(11):1183－1191.
[20] 杨卫红,赖清平,兰宇,等. 基于调节潜力指标的用户用电行为聚类分析算法研究[J]. 电力建设,2018,39(6):96－104.
[21] 陈娟,吴卓青,邓胜利. 基于层次聚类法的"知乎"用户细分与行为分析[J]. 情报理论与实践,2018,41(7):111－116.
[22] MOHAMMED E A, WESAM M. Efficient and fast initialization algorithm for K-means clustering[J]. Intelligent systems and applications, 2012,4(1):21－31.
[23] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492－1496.
[24] 杨洁,王国胤,庞紫玲. 密度峰值聚类相关问题的研究[J]. 南京大学学报(自然科学),2017,53(4):791－801.
[25] 亓丛,吴俊. 用户画像概念溯源与应用场景研究[J]. 重庆交通大学学报(社会科学版),2017,17(5):82－87.
[26] 杨磊. 基于改进RFM模型的聚类算法在农村用户4G消费行为中研究与应用[D]. 南京:南京邮电大学,2017.
[27] 伍育红. 聚类算法综述[J]. 计算机科学, 2015(S1): 491－499,524.
[28] 王洋,张桂珠. 自动确定聚类中心的密度峰值算法[J]. 计算机工程与应用,2018(8):137－142.
[29] 卢建云,朱庆生,吴全旺. 一种启发式确定聚类数方法[J]. 小型微型计算机系统,2018,39(7):1381－1385.
[30] 魏红燕,魏含玉. 基于密度峰值聚类的用户类型划分[J]. 商丘师范学院学报,2017(12):11－13.

作者贡献说明:
张莉曼:负责论文撰写、数据处理分析;
张向先:负责论文框架的指导与确定;
卢恒:负责数据处理;
张玲:负责论文最后审阅及定稿。

Research on User Persona of Knowledge Online Live's Paid-Up Members

Zhang Liman　Zhang Xiangxian　Lu Heng　Zhang Ling
School of Management, Jilin University, Changchun 130022

Abstract: [**Purpose/significance**] This paper aims at giving effective enlightenment and method for the platform to locate the characteristics of the users, carry out targeted operations and discover the potential value users. [**Method/process**] This paper takes the knowledge live platform as the research object, firstly designing the concept model of the group portrait, then putting forward the method of the paid-up user persona of the knowledge live platform based on the density peak clustering, and finally taking the data of paid-up users in Zhihu live platform as an sample, and using the density peak clustering algorithm in Matlab to divide the sample data into groups. [**Result/conclusion**] The experimental results show that there are five typical user groups in Zhihu live platform. By comparing and analyzing the cluster center points, we identify the typical characteristics of each group, and propose the precise operation strategy.

Keywords: knowledge online live　density peaks clustering　user persona　paid-up members